Chapter 1

A Short History of Markov Chain Monte Carlo

Subjective Recollections from Incomplete Data

Christian Robert[1]
Université Paris Dauphine and CREST, INSEE
and *George Casella*[2]
University of Florida

1.1 Introduction

Markov chain Monte Carlo (MCMC) methods have been around for almost as long as Monte Carlo techniques, even though their impact on statistics has not been truly felt until the very early 1990s, except in the specialized fields of spatial statistics and image analysis, where those methods appeared earlier. The emergence of Markov based techniques in physics is a story that remains un-

[1]Professor of Statistics, Université Paris-Dauphine, CEREMADE, 75775 Paris cedex 16, France. Supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2006-2008 project ANR=05-BLAN-0299 Adap'MC. This work was partly done during a visit to the Queensland University of Technology, Brisbane, and the author is grateful to Kerrie Mengersen for her hospitality and support. Email: xian@ceremade.dauphine.fr.

[2] Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grants DMS-04-05543, DMS-0631632 and SES-0631588. Email: casella@ufl.edu.

told within this survey (see Landau and Binder 2005). Also, we will not enter into a description of MCMC techniques, unless they have some historical link, as the remainder of this volume covers the technical aspects. A comprehensive treatment with further references can also be found in Robert and Casella (2004).

We will distinguish between the introduction of Metropolis-Hastings based algorithms and those related to Gibbs sampling, since they each stem from radically different origins, even though their mathematical justification via Markov chain theory is the same. Tracing the development of Monte Carlo methods, we will also briefly mention what we might call the "second-generation MCMC revolution". Starting in the mid-to-late 1990s, this includes the development of particle filters, reversible jump and perfect sampling, and concludes with more current work on population or sequential Monte Carlo and regeneration and the computing of "honest" standard errors.

As mentioned above, the realization that Markov chains could be used in a wide variety of situations only came (to mainstream statisticians) with Gelfand and Smith (1990), despite earlier publications in the statistical literature like Hastings (1970), Geman and Geman (1984) and Tanner and Wong (1987). Several reasons can be advanced: lack of computing machinery (think of the computers of 1970!), or background on Markov chains, or hesitation to trust in the practicality of the method. It thus required visionary researchers like Gelfand and Smith to convince the community, supported by papers that demonstrated, through a series of applications, that the method was easy to understand, easy to implement and practical (Gelfand et al. 1990, 1992, Smith and Gelfand 1992, Wakefield et al. 1994). The rapid emergence of the dedicated BUGS (Bayesian inference Using Gibbs Sampling) software as early as 1991, when a paper on BUGS was presented at the Valencia meeting, was another compelling argument for adopting, at large, MCMC algorithms.[3]

1.2 Before the Revolution

Monte Carlo methods were born in Los Alamos, New Mexico during World War II, eventually resulting in the Metropolis algorithm in the early 1950s. While Monte Carlo methods were in use by that time, MCMC was brought closer to

[3] Historically speaking, the development of BUGS initiated from Geman and Geman (1984) and Pearl (1987), in accord with the developments in the artificial intelligence community, and it pre-dates Gelfand and Smith (1990).

statistical practicality by the work of Hastings in the 1970s.

What can be reasonably seen as the first MCMC algorithm is what we now call the Metropolis algorithm, published by Metropolis et al. (1953). It emanates from the same group of scientists who produced the Monte Carlo method, namely the research scientists of Los Alamos, mostly physicists working on mathematical physics and the atomic bomb.

MCMC algorithms therefore date back to the same time as the development of regular (MC only) Monte Carlo methods, which are usually traced to Ulam and von Neumann in the late 1940s. Stanislaw Ulam associates the original idea with an intractable combinatorial computation he attempted in 1946 (calculating the probability of winning at the card game "solitaire"). This idea was enthusiastically adopted by John von Neumann for implementation with direct applications to neutron diffusion, the name "Monte Carlo" being suggested by Nicholas Metropolis. (Eckhardt 1987 describes these early Monte Carlo developments, and Hitchcock 2003 gives a brief history of the Metropolis algorithm.)

These occurrences very closely coincide with the appearance of the very first general-purpose digital computer, the ENIAC, which came to life in February 1946, after three years of construction. The Monte Carlo method was set up by von Neumann, who was using it on thermonuclear and fission problems as early as 1947. At the same time, that is, 1947, Ulam and von Neumann invented inversion and accept-reject techniques (also recounted in Eckhardt 1987) to simulate from non-uniform distributions. Without computers, a rudimentary version invented by Fermi in the 1930s did not get any recognition (Metropolis 1987). Note also that, as early as 1949, a symposium on Monte Carlo was supported by Rand, NBS and the Oak Ridge laboratory and that Metropolis and Ulam (1949) published the very first paper about the Monte Carlo method.

1.2.1 The Metropolis et al. (1953) paper

The first MCMC algorithm is associated with a second computer, called MANIAC,[4], built in Los Alamos under the direction of Metropolis in early 1952. Both a physicist and a mathematician, Nicholas Metropolis, who died in Los Alamos in 1999, came to this place in April 1943 . The other members of the team also came to Los Alamos during those years, including the controversial Edward Teller. As early as 1942, this physicist became obsessed with the hydrogen (H) bomb, which he eventually managed to design with Stanislaw Ulam,

[4]MANIAC stands for *Mathematical Analyzer, Numerical Integrator and Computer.*

using the better computer facilities in the early 1950s.

Published in June 1953 in the *Journal of Chemical Physics*, the primary focus of Metropolis et al. (1953) is the computation of integrals of the form

$$\Im = \int F(\theta) \exp\{-E(\theta)/kT\} \mathrm{d}\theta \Big/ \int \exp\{-E(\theta)/kT\} \mathrm{d}\theta \,,$$

on $\mathbb{R}^{2N}$, θ denoting a set of N particles on $\mathbb{R}^2$, with the energy E being defined as

$$E(\theta) = \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} V(d_{ij}),$$

where V a potential function and d_{ij} the Euclidean distance between particles i and j in θ. The *Boltzmann distribution* $\exp\{-E(\theta)/kT\}$ is parameterized by the *temperature* T, k being the Boltzmann constant, with a normalization factor

$$Z(T) = \int \exp\{-E(\theta)/kT\} \mathrm{d}\theta,$$

that is not available in closed form, except in trivial cases. Since θ is a $2N$-dimensional vector, numerical integration is impossible. Given the large dimension of the problem, even standard Monte Carlo techniques fail to correctly approximate $\Im$, since $\exp\{-E(\theta)/kT\}$ is very small for most realizations of the random configurations of the particle system (uniformly in the $2N$ square). In order to improve the efficiency of the Monte Carlo method, Metropolis et al. (1953) propose a random walk modification of the N particles. That is, for each particle i ($1 \leq i \leq N$), values

$$x_i' = x_i + \sigma\xi_{1i} \quad \text{and} \quad y_i' = y_i + \sigma\xi_{2i}$$

are proposed, where both ξ_{1i} and ξ_{2i} are uniform $\mathcal{U}(-1, 1)$. The energy difference ΔE between the new configuration and the previous one is then computed and the new configuration is accepted with probability

$$\min\{1, \exp(-\Delta E/kT)\} \,, \tag{1.1}$$

and otherwise the previous configuration is replicated, in the sense that its counter is increased by one in the final average of the $F(\theta_t)$'s over the τ moves of the random walk, $1 \leq t \leq \tau$). Note that Metropolis et al. (1953) move one

particle at a time, rather than moving all of them together, which makes the initial algorithm appear as a primitive kind of Gibbs sampler!

The authors of Metropolis et al. (1953) demonstrate the validity of the algorithm by first establishing irreducibility, which they call *ergodicity*, and second proving ergodicity, that is, convergence to the stationary distribution. The second part is obtained via a discretization of the space: They first note that the proposal move is reversible, then establish that $\exp\{-E/kT\}$ is invariant. The result is therefore proven in its full generality, minus the discretization. The number of iterations of the Metropolis algorithm used in the paper seems to be limited: 16 steps for burn-in and 48 to 64 subsequent iterations, which required four to five hours on the Los Alamos computer.

An interesting variation is the *Simulated Annealing* algorithm, developed by Kirkpatrick et al. (1983), who connected optimization with *annealing*, the cooling of a metal. Their variation is to allow the temperature T in (1.1) to decrease as the algorithm runs, according to a "cooling schedule". The Simulated Annealing algorithm can be shown to find the global maximum with probability 1, although the analysis is quite complex due to the fact that, with varying T, the algorithm is no longer a time-homogeneous Markov chain.

1.2.2 The Hastings (1970) paper

The Metropolis algorithm was later generalized by Hastings (1970) and his student Peskun (1973, 1981) as a statistical simulation tool that could overcome the curse of dimensionality met by regular Monte Carlo methods, a point already emphasized in Metropolis et al. 1953.[5]

In his *Biometrika* paper,[6] Hastings (1970) also defines his methodology for finite and reversible Markov chains, treating the continuous case by using a discretization analogy. The generic probability of acceptance for a move from state i to state j is

$$\alpha_{ij} = \frac{s_{ij}}{1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}}\,,$$

where $s_{ij} = s_{ji}$ is a positive quantity ensuring that $\alpha_{ij} \leq 1$, π_i denotes the target and q_{ij} the proposal. This generic form of probability encompasses the

[5] In fact, Hastings starts by mentioning a decomposition of the target distribution into a *product of one-dimensional conditional distributions* but this falls short of an early Gibbs sampler.

[6] Hastings (1970) is one of the ten *Biometrika* papers reproduced in Titterington and Cox (2001).

forms of both Metropolis et al. (1953) and Barker (1965). At this stage, Hastings mentions that *little is known about the relative merits of those two choices* (even though) *Metropolis's method may be preferable*. He also warns against *high rejection rates as indicative of a poor choice of transition matrix*, but does not mention the opposite pitfall of low rejection rates, associated with a slow exploration of the target.

The examples in the paper include a Poisson target with a ± 1 random walk proposal and a normal target with a uniform random walk proposal mixed with its reflection, i.e. a uniform proposal centered at $-\theta_t$ rather than at the current value θ_t of the Markov chain. On a multivariate target, Hastings introduces a Gibbs sampling strategy, updating one component at a time and defining the composed transition as satisfying the stationary condition because each component does leave the target invariant. Hastings (1970) actually refers to Erhman et al. (1960) as a preliminary, if specific, instance of this sampler. More precisely, this is Metropolis-within-Gibbs except for the name. This first introduction of the Gibbs sampler has thus been completely overlooked, even though the proof of convergence is completely general, based on a composition argument as in Tierney (1994), discussed in Section 1.4.1. The remainder of the paper deals with (a) an importance sampling version of MCMC, (b) general remarks about assessment of the error, and (c) an application to random orthogonal matrices, with another example of Gibbs sampling.

Three years later, Peskun (1973) published a comparison of Metropolis' and Barker's forms of acceptance probabilities and showed in a discrete setup that the optimal choice is that of Metropolis, where optimality is to be understood in terms of the asymptotic variance of any empirical average. The proof is a direct consequence of a result by Kemeny and Snell (1960) on the asymptotic variance. Peskun also establishes that this asymptotic variance can improve upon the iid case if and only if the eigenvalues of $\mathbf{P} - \mathbf{A}$ are all negative, when $\mathbf{A}$ is the transition matrix corresponding to iid simulation and $\mathbf{P}$ the transition matrix corresponding to the Metropolis algorithm, but he concludes that the trace of $\mathbf{P} - \mathbf{A}$ is always positive, therefore that the uniform improvement is impossible.

1.3 Seeds of the Revolution

A number of early pioneers had brought forward the seeds of Gibbs sampling; in particular, Hammersley and Clifford had produced a constructive argument in 1970 to recover a joint distribution from its conditionals, a result later called the *Hammersley–Clifford* theorem by Besag (1974, 1986). Besides Hastings (1970) and Geman and Geman (1984), already mentioned, other papers that contained the seeds of Gibbs sampling are Besag and Clifford (1989), Broniatowski et al. (1984), Qian and Titterington (1990), and Tanner and Wong (1987).

1.3.1 Besag and the Fundamental (Missing) Theorem

In the early 1970's, Hammersley, Clifford, and Besag were working on the specification of joint distributions from conditional distributions and on necessary and sufficient conditions for the conditional distributions to be compatible with a joint distribution. What is now known as the *Hammersley-Clifford* theorem states that a joint distribution for a vector associated with a dependence graph (edge meaning dependence and absence of edge conditional independence) must be represented as a product of functions over the *cliques* of the graphs, that is, of functions depending only on the components indexed by the labels in the clique[7].

From an historical point of view, Hammersley (1974) explains why the Hammersley-Clifford theorem was never published as such, but only through Besag (1974). The reason is that Clifford and Hammersley were dissatisfied with the positivity constraint: The joint density could be recovered from the full conditionals only when the support of the joint was made of the product of the supports of the full conditionals. While they strived *to make the theorem independent of any positivity condition*, their graduate student published a counter-example that put a full stop to their endeavors (Moussouris 1974).

While Besag (1974) can certainly be credited to some extent of the (re-)discovery of the Gibbs sampler, Besag (1975) expressed doubt about the practicality of his method, noting that "the technique is unlikely to be particularly helpful in many other than binary situations and the Markov chain itself has no practical interpretation", clearly understating the importance of his work.

A more optimistic sentiment was expressed earlier by Hammersley and Handscomb (1964), in their textbook on Monte Carlo methods. There they cover such

[7] A clique is a maximal subset of the nodes of a graphs such that every pair of nodes within the clique is connected by an edge (Cressie 1993).

topics as "Crude Monte Carlo"; importance sampling; control variates; and "Conditional Monte Carlo", which looks surprisingly like a simulation approach to missing-data models (see Section 1.3.2). Of course, they do not cover the Hammersley-Clifford theorem but they state in the Preface: *"We are convinced nevertheless that Monte Carlo methods will one day reach an impressive maturity."* Well said!

1.3.2 EM and its Simulated Versions as Precursors

Due to its connection with missing data problems, the EM algorithm (Dempster et al. 1977) has early connections with Gibbs sampling.[8] For instance, Broniatowski et al. (1984) and Celeux and Diebolt (1985) had tried to overcome the dependence of EM methods on the starting value by replacing the E step with a *simulation* step, the missing data z_m being generated conditionally on the observation x and on the current value of the parameter θ_m. The maximization in the M step is then operated on the simulated complete-data likelihood, $L(\theta|x, z_m)$, producing a new value θ_{m+1} and this appears as a predecessor to the Gibbs step of Gelman and King (1990) and Diebolt and Robert (1994) for mixture estimation.[9] Unfortunately, the theoretical convergence results for these methods are limited. Celeux and Diebolt (1990) have, however, solved the convergence problem of SEM by devising a hybrid version called SAEM (for *Simulated Annealing EM*), where the amount of randomness in the simulations decreases with the iterations, ending up with an EM algorithm.[10]

1.3.3 Gibbs, and Beyond

Although somewhat removed from statistical inference in the classical sense and based on earlier techniques used in statistical physics, the landmark paper by Geman and Geman (1984) brought Gibbs sampling into the arena of statistical application. This paper is also responsible for the name *Gibbs sampling*, because it implemented this method for the Bayesian study of *Gibbs random fields* which,

[8] This is especially relevant when considering the early introduction of a Gibbs sampler by data augmentation in Tanner and Wong (1987).

[9] The achievement in the former paper remained unoticed for several years due to the low-key and off-handed use of the Gibbs sampler at a time when it was unknown to most of the community.

[10] Other and more well-known connections between EM and MCMC algorithms can be found in the literature (Liu and Rubin 1994, Meng and Rubin 1992, Wei and Tanner 1990), but the connection with Gibbs sampling is more tenuous in that the simulation methods there are used to approximate quantities in a Monte Carlo fashion.

in turn, derive their name from the physicist Josiah Willard Gibbs (1839–1903). This original implementation of the Gibbs sampler was applied to a discrete image processing problem and did not involve completion as in Section 1.3.2. But this was one more spark that led to the explosion, as it had a clear influence on Green, Smith, Spiegelhalter and others.

The extent to which Gibbs sampling and Metropolis algorithms were in use within the image analysis and point process communities is actually quite large, as illustrated in Ripley (1987) where Section §4.7 is entitled "Metropolis' method and random fields" and describes the implementation and the validation of the Metropolis algorithm in a finite setting with an application to Markov random fields and the corresponding issue of bypassing the normalizing constant. Besag et al. (1991) is another striking example of the activity in the spatial statistics community at the end of the 1980's.

1.4 The Revolution

The gap of more than 30 years between Metropolis et al. (1953) and Gelfand and Smith (1990) can still be partially attributed to the lack of appropriate computing power, as most of the examples now processed by MCMC algorithms could not have been treated previously, even though the hundreds of dimensions processed in Metropolis et al. (1953) were quite formidable. However, by the mid-1980s, the pieces were all in place.

After Peskun, MCMC in the statistical world was dormant for about 10 years, and then several papers appeared that highlighted its usefulness in specific settings like pattern recognition, image analysis or spatial statistics. In particular, Geman and Geman (1984) influenced Gelfand and Smith (1990) to write a paper that is the genuine starting point for an intensive use of MCMC methods by the mainstream statistical community. It sparked new interest in Bayesian methods, statistical computing, algorithms, and stochastic processes through the use of computing algorithms such as the Gibbs sampler and the Metropolis–Hastings algorithm. Casella and George 1992 wrote an elementary introduction to the Gibbs sampler[11] in *American Statistician* that disseminated the technique to a wider community while explaining in simple terms why the

[11] On a humorous note, the original Technical Report of this paper was called *Gibbs for Kids*, which was changed because a referee did not appreciate the humor. However, our colleague Dan Gianola, an Animal Breeder at Wisconsin, liked the title. In using Gibbs sampling in his work, he gave a presentation in 1993 at the 44th Annual Meeting of the European Association for Animal Production, Arhus, Denmark. The title: *Gibbs for Pigs*.

algorithm is valid.

Interestingly, the earlier paper by Tanner and Wong (1987) had essentially the same ingredients as Gelfand and Smith (1990), namely the fact that simulating from the conditional distributions is sufficient to asymptotically simulate from the joint. This paper was considered important enough to be a discussion paper in the *Journal of the American Statistical Association*, but its impact was somehow limited, compared with Gelfand and Smith (1990). There are several reasons for this; one being that the method seemed to only apply to missing data problems, this impression being reinforced by the name *data augmentation*, and another is that the authors were more focused on approximating the posterior distribution. They suggested a MCMC approximation to the target $\pi(\theta|x)$ at each iteration of the sampler, based on

$$\frac{1}{m}\sum_{k=1}^{m}\pi(\theta|x,z^{t,k})\,, \qquad z^{t,k}\sim\hat{\pi}_{t-1}(z|x)\,, \quad k=1,\ldots,m,$$

that is, by replicating m times the simulations from the current approximation $\hat{\pi}_{t-1}(z|x)$ of the marginal posterior distribution of the missing data. This focus on estimation of the posterior distribution connected the original Data Augmentation algorithm to EM, as pointed out by Dempster in the discussion. Although the discussion by Morris gets very close to the two-stage Gibbs sampler for hierarchical models, he is still concerned about doing m iterations, and worries about how costly that would be. Tanner and Wong mention taking $m = 1$ at the end of the paper, referring to this as an "extreme case".

In a sense, Tanner and Wong (1987) was still too close to Rubin's 1978 multiple imputation to start a new revolution. Yet another reason for this may be that the theoretical background was based on functional analysis rather than Markov chain theory, which needed, in particular, for the Markov kernel to be uniformly bounded and equicontinuous. This may have discouraged potential users as requiring too much mathematics.

The authors of this review were fortunate enough to attend many focused conferences during this time, where we were able to witness the explosion of Gibbs sampling. In the summer of 1986 in Bowling Green, Ohio, Smith gave a series of ten lectures on hierarchical models. Although there was a lot of computing mentioned, the Gibbs sampler was not fully developed yet. (Interestingly, Smith commented that the limiting factor, at that time, for the full exploitation of hierarchical models in statistical problem, was the inability to compute high-

dimensional integrals.) In another lecture in June 1989 at a Bayesian workshop in Sherbrooke, Québec, he revealed for the first time the generic features of Gibbs sampler, and we still remember vividly the shock induced on ourselves and on the whole audience by the sheer breadth of the method: This development of Gibbs sampling, MCMC, and the resulting seminal paper of Gelfand and Smith (1990) was an *epiphany* in the world of statistics.

Definition: `epiphany` *n.* A spiritual event in which the essence of a given object of manifestation appears to the subject, as in a sudden flash of recognition.

The explosion had begun, and just two years later, an MCMC conference at Ohio State University organized by Gelfand, Goel, and Smith, consisted of three full days of talks. Many of the talks were to become influential papers; including Albert and Chib (1993), Gelman and Rubin (1992), Geyer (1992), Gilks (1992), Liu et al. (1994, 1995), and Tierney (1994).

Approximately one year later, in May of 1992, there was a meeting of the Royal Statistical Society on "The Gibbs sampler and other Markov chain Monte Carlo methods", where four papers were presented followed by much discussion. The papers appear in the first volume of JRSSB in 1993, together with 49 (!) pages of discussion. The excitement is clearly evident in the writings, even though the theory and implementation were not always perfectly understood.

Looking at these meetings, we can see the paths that Gibbs sampling would lead us down. In the next two sections we will summarize some of the advances from the early to mid 1990s.

1.4.1 Advances in MCMC Theory

Perhaps the most influential MCMC theory paper of the 1990s is Tierney (1994), who carefully laid out all of the assumptions needed to analyze the Markov chains and then developed their properties, in particular, convergence of ergodic averages and central limit theorems. In one of the discussions of that paper, Chan and Geyer (1994) were able to relax a condition on Tierney's Central Limit Theorem, and this new condition plays an important role in research today (see Section 1.5.4). A pair of very influential, and innovative, papers is the work of Liu et al. (1994, 1995), who very carefully analyzed the covariance structure of Gibbs sampling, and were able to formally establish the validity of Rao-Blackwellization in Gibbs sampling. Gelfand and Smith (1990) had used Rao-Blackwellization, but it was not justified at that time, as the original theorem

was only applicable to iid sampling, which is not the case in MCMC. Another significant entry is Rosenthal (1995), who obtained one of the earliest results on exact rates of convergence.

Another paper must be singled out, namely Mengersen and Tweedie (1996), for setting the tone for the study of the speed of convergence of MCMC algorithms to the target distribution. Subsequent works in this area by Richard Tweedie, Gareth Roberts, Jeff Rosenthal and co-authors are too numerous to be mentioned here, even though the paper by Roberts et al. (1997) must be cited for setting explicit targets on the acceptance rate of the random walk Metropolis–Hastings algorithm, as well as Roberts and Rosenthal (1999) for getting an upper bound on the number of iterations (523) needed to approximate the target up to 1% by a slice sampler. The untimely death of Richard Tweedie in 2001 alas had a major impact on the book about MCMC convergence he was contemplating with Gareth Roberts.

One pitfall arising from the widespread use of Gibbs sampling was the tendency to specify models only through their conditional distributions, almost always without referring to the positivity conditions in Section 1.3. Unfortunately, it is possible to specify a perfectly legitimate-looking set of conditionals that do not correspond to any joint distribution, and the resulting Gibbs chain cannot converge. Hobert and Casella (1996) were able to document the conditions needed for a convergent Gibbs chain, and alerted the Gibbs community to this problem, which only arises when improper priors are used, but this is a frequent occurrence.

Much other work followed, and continues to grow today. Geyer and Thompson (1995) describe how to put a "ladder" of chains together to have both "hot" and "cold" exploration, followed by Neal's 1996 introduction of tempering; Athreya et al. (1996) gave more easily verifiable conditions for convergence; Meng and van Dyk (1999) and Liu and Wu (1999) developed the theory of parameter expansion in the Data Augmentation algorithm, leading to construction of chains with faster convergence, and to the work of Hobert and Marchev (2008), who give precise constructions and theorems to show how parameter expansion can uniformly improve over the original chain.

1.4.2 Advances in MCMC Applications

The real reason for the explosion of MCMC methods was the fact that an enormous number of problems that were deemed to be computational nightmares

now cracked open like eggs. As an example, consider this very simple random effects model from Gelfand and Smith (1990). Observe

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, J, \tag{1.2}$$

where

$$\begin{aligned} \theta_i &\sim \mathrm{N}(\mu, \sigma_\theta^2) \\ \varepsilon_{ij} &\sim \mathrm{N}(0, \sigma_\varepsilon^2), \text{ independent of } \theta_i \end{aligned}$$

Estimation of the variance components can be difficult for a frequentist (REML is typically preferred) but it indeed was a nightmare for a Bayesian, as the integrals were intractable. However, with the usual priors on μ, σ_θ^2, and σ_ε^2, the full conditionals are trivial to sample from and the problem is easily solved via Gibbs sampling. Moreover, we can increase the number of variance components and the Gibbs solution remains easy to implement.

During the early 1990s, researchers found that Gibbs, or Metropolis-Hastings, algorithms would crack almost any problem that they looked at, and there was a veritable flood of papers applying MCMC to previously intractable models, and getting good solutions. For example, building on (1.2), it was quickly realized that Gibbs sampling was an easy route to getting estimates in the linear mixed models (Wang et al. 1993, 1994), and even generalized linear mixed models (Zeger and Karim 1991). Building on the experience gained with the EM algorithm, similar arguments made it possible to analyze probit models using a latent variable approach in a linear mixed model (Albert and Chib 1993), and in mixture models with Gibbs sampling (Diebolt and Robert 1994). It progressively dawned on the community that latent variables could be artificially introduced to run the Gibbs sampler in about every situation, as eventually published in Damien et al. (1999), the main example being the slice sampler (Neal 2003). A very incomplete list of some other applications include changepoint analysis (Carlin et al. 1992, Stephens 1994); Genomics (Churchill 1995, Lawrence et al. 1993, Stephens and Smith 1993); capture-recapture (Dupuis 1995, George and Robert 1992); variable selection in regression (George and McCulloch 1993); spatial statistics (Raftery and Banfield 1991), and longitudinal studies (Lange et al. 1992).

Many of these applications were advanced though other developments such as the Adaptive Rejection Sampling of Gilks (1992), Gilks et al. (1995), and the

simulated tempering approaches of Geyer and Thompson (1995) or Neal (1996).

1.5 After the Revolution

After the revolution comes the "second" revolution, but now we have a more mature field. The revolution has slowed, and the problems are being solved in, perhaps, deeper and more sophisticated ways, even though Gibbs sampling also offers to the amateur the possibility to handle Bayesian analysis in complex models at little cost, as exhibited by the widespread use of BUGS, which mostly focuses[12] on this approach. But, as before, the methodology continues to expand the set of problems for which statisticians can provide meaningful solutions, and thus continues to further the impact of statistics.

1.5.1 A Brief Glimpse at Particle Systems

The realization of the possibilities of iterating importance sampling is not new: in fact, it is about as old as Monte Carlo methods themselves. It can be found in the molecular simulation literature of the 50's, as in Hammersley and Morton (1954), Rosenbluth and Rosenbluth (1955) and Marshall (1965). Hammersley and colleagues proposed such a method to simulate a self-avoiding random walk (see Madras and Slade 1993) on a grid, due to huge inefficiencies in regular importance sampling and rejection techniques. Although this early implementation occurred in particle physics, the use of the term "particle" only dates back to Kitagawa (1996), while Carpenter et al. (1997) coined the term "particle filter". In signal processing, early occurrences of a particle filter can be traced back to Handschin and Mayne (1969).

More in connection with our theme, the landmark paper of Gordon et al. (1993) introduced the bootstrap filter which, while formally connected with importance sampling, involves past simulations and possible MCMC steps (Gilks and Berzuini 2001). As described in the volume edited by Doucet et al. (2001), particle filters are simulation methods adapted to sequential settings where data are collected progressively in time as in radar detection, telecommunication correction or financial volatility estimation. Taking advantage of state-space representations of those dynamic models, particle filter methods produce Monte Carlo approximations to the posterior distributions by propagating simulated samples whose weights are actualized against the incoming observations. Since the importance

[12] BUGS now uses both Gibbs sampling and Metropolis-Hastings algorithms.

weights have a tendency to degenerate, that is, all weights but one are close to zero, additional MCMC steps can be introduced at times to recover the variety and representativeness of the sample. Modern connections with MCMC in the construction of the proposal kernel are to be found, for instance, in Doucet et al. (2000) and in Del Moral et al. (2006). In parallel, sequential imputation was developed in Kong et al. (1994), while Liu and Chen (1995) first formally pointed out the importance of resampling in sequential Monte Carlo, a term coined by them.

The recent literature on the topic more closely bridges the gap between sequential Monte Carlo and MCMC methods by making adaptive MCMC a possibility (see, for example, Andrieu et al. 2004 or Roberts and Rosenthal 2005).

1.5.2 Perfect sampling

Introduced in the seminal paper of Propp and Wilson (1996), perfect sampling, namely the ability to use MCMC methods to produce an exact (or perfect) simulation from the target, maintains a unique place in the history of MCMC methods. Although this exciting discovery led to an outburst of papers, in particular in the large body of work of Møller and coauthors, including the book by Møller and Waagepetersen (2003), as well as many reviews and introductory materials, like Casella et al. (2001), Fismen (1998), and Dimakos (2001), the excitement quickly dried out. The major reason for this ephemeral lifespan is that the construction of perfect samplers is most often close to impossible or impractical, despite some advances in the implementation (Fill 1998a,b).

There is, however, ongoing activity in the area of point processes and stochastic geometry, much from the work of Møller and Kendall. In particular, Kendall and Møller (2000) developed an alternative to the *Coupling From The Past* (CFTP) algorithm of Propp and Wilson (1996), called *horizontal CFTP*, which mainly applies to point processes and is based on continuous time birth-and-death processes. See also Fernández et al. (1999) for another horizontal CFTP algorithm for point processes. Berthelsen and Møller (2003) exhibited a use of these algorithms for nonparametric Bayesian inference on point processes.

1.5.3 Reversible jump and variable dimensions

From many viewpoints, the invention of the reversible jump algorithm in Green (1995) can be seen as the start of the second MCMC revolution: the formalization of a Markov chain that moves across models and parameter spaces al-

lowed for the Bayesian processing of a wide variety of new models and contributed to the success of Bayesian model choice and subsequently to its adoption in other fields. There exist earlier alternative Monte Carlo solutions like Gelfand and Dey (1994) and Carlin and Chib (1995), the later being very close in spirit to reversible jump MCMC (as shown by the completion scheme of Brooks et al. 2003), but the definition of a proper balance condition on cross-model Markov kernels in Green (1995) gives a generic setup for exploring variable dimension spaces, even when the number of models under comparison is infinite. The impact of this new idea was clearly perceived when looking at the First European Conference on Highly Structured Stochastic Systems that took place in Rebild, Denmark, the next year, organized by Stephen Lauritzen and Jesper Møller: a large majority of the talks were aimed at direct implementations of RJMCMC to various inference problems. The application of RJMCMC to mixture order estimation in the discussion paper of Richardson and Green (1997) ensured further dissemination of the technique. More recently, Stephens (2000) proposed a continuous time version of RJMCMC, based on earlier ideas of Geyer and Møller (1994), but with similar properties (Cappé et al. 2003), while Brooks et al. (2003) made proposals for increasing the efficiency of the moves. In retrospect, while reversible jump is somehow unavoidable in the processing of very large numbers of models under comparison, as for instance in variable selection (Marin and Robert 2007), the implementation of a complex algorithm like RJMCMC for the comparison of a few models is somewhat of an overkill since there exist alternative solutions based on model specific MCMC chains, for example (Chen et al. 2000).

1.5.4 Regeneration and the CLT

While the Central Limit Theorem (CLT) is a central tool in Monte Carlo convergence assessment, its use in MCMC setups took longer to emerge, despite early signals by Geyer (1992), and it is only recently that sufficiently clear conditions emerged. We recall that the Ergodic Theorem (see, for example, Robert and Casella 2004, Theorem 6.63) states that, if $(\theta_t)_t$ is a Markov chain with stationary distribution π, and $h(\cdot)$ is a function with finite variance, then under fairly mild conditions,

$$\lim_{n\to\infty} \bar{h}_n = \int h(\theta)\pi(\theta)\, d\theta = \mathrm{E}_\pi[h(\theta)], \tag{1.3}$$

almost everywhere, where $\bar{h}_n = (1/n)\sum_{i=1}^{n} h(\theta_i)$. For the CLT to be used to monitor this convergence,

$$\frac{\sqrt{n}(\bar{h}_n - \mathrm{E}_\pi[h(\theta)])}{\sqrt{\mathrm{Var}[h(\theta)]}} \to \mathrm{N}(0,1), \tag{1.4}$$

there are two roadblocks. First, convergence to normality is strongly affected by the lack of independence. To get CLTs for Markov chains, we can use a result of Kipnis and Varadhan (1986), which requires the chain to be reversible, as is the case for Metropolis-Hastings chains, or we must delve into mixing conditions (Billingsley 1995, Section 27), which are typically not easy to verify. However, Chan and Geyer (1994) showed how the condition of geometric ergodicity could be used to establish CLTs for Markov chains. But getting the convergence is only half of the problem. In order to use (1.4), we must be able to consistently estimate the variance, which turns out to be another difficult endeavor. The "naïve" estimate of the usual standard error is not consistent in the dependent case and the most promising paths for consistent variance estimates seems to be through regeneration and batch means.

The theory of regeneration uses the concept of a split chain (Athreya and Ney 1978), and allows us to independently restart the chain while preserving the stationary distribution. These independent "tours" then allow the calculation of consistent variance estimates and honest monitoring of convergence through (1.4). Early work on applying regeneration to MCMC chains was done by Mykland et al. (1995) and Robert (1995), who showed how to construct the chains and use them for variance calculations and diagnostics (see also Guihenneuc-Jouyaux and Robert 1998), as well as deriving adaptive MCMC algorithms (Gilks et al. 1998). Rosenthal (1995) also showed how to construct and use regenerative chains, and much of this work is reviewed in Jones and Hobert (2001). The most interesting and practical developments, however, are in Hobert et al. (2002) and Jones et al. (2006), where consistent estimators are constructed for $\mathrm{Var}[h(\theta)]$, allowing valid monitoring of convergence in chains that satisfy the CLT. Interestingly, although Hobert et al. (2002) uses regeneration, Jones et al. (2006) get their consistent estimators thorough another technique, that of cumulative batch means.

1.6 Conclusion

The impact of Gibbs sampling and MCMC on Bayesian statistics was to change our entire method of thinking and attacking problems, representing a *paradigm shift* (Kuhn 1996). Now, the collection of real problems that we could solve grew almost without bound. Markov chain Monte Carlo changed our emphasis from "closed form" solutions to algorithms, expanded our impact to solving "real" applied problems and to improving numerical algorithms using statistical ideas, and led us into a world where "exact" now means "simulated".

This has truly been a quantum leap in the evolution of the field of statistics, and the evidence is that there are no signs of slowing down. Although the "explosion" is over, the current work is going deeper into theory and applications, and continues to expand our horizons and influence by increasing our ability to solve even bigger and more important problems. The size of the data sets, and of the models, for example in genomics or climatology, is something that could not have been conceived 60 years ago, when Ulam and von Neumann invented the Monte Carlo method. Now we continue to plod on, and hope that the advances that we make here will, in some way, help our colleagues 60 years in the future solve problems that we cannot yet conceive.

Acknowledgement: We are grateful for comments and suggestions from Olivier Cappé, Alan Gelfand, Peter Green, Jun Liu, Sharon McGrayne, Peter Müller, Gareth Roberts, Adrian Smith, and David Spiegelhalter.

Bibliography

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. American Statist. Assoc.*, 88:669–679.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. (2004). An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43.

Athreya, K., Doss, H., and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *Ann. Statist.*, 24:69–100.

Athreya, K. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.*, 245:493–501.

Barker, A. (1965). Monte Carlo calculations of the radial distribution functions for a proton electron plasma. *Aust. J. Physics*, 18:119–133.

Berthelsen, K. and Møller, J. (2003). Likelihood and non-parametric Bayesian MCMC inference for spatial point processes based on perfect simulation and path sampling. *Scandinavian J. Statist.*, 30:549–564.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Society Series B*, 36:192–326. With discussion.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195.

Besag, J. (1986). On the statistical analysis of dirty pictures. *J. Royal Statist. Society Series B*, 48:259–279.

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, 76:633–642.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals Inst. Statist. Mathematics*, 42(1):1–59.

Billingsley, P. (1995). *Probability and Measure*. John Wiley, New York, third edition.

Broniatowski, M., Celeux, G., and Diebolt, J. (1984). Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. In Diday, E., editor, *Data Analysis and Informatics*, volume 3, pages 359–373. North-Holland, Amsterdam.

Brooks, S., Giudici, P., and Roberts, G. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. Royal Statist. Society Series B*, 65(1):3–55.

Cappé, O., Robert, C., and Rydén, T. (2003). Reversible jump, birth-and-death, and more general continuous time MCMC samplers. *J. Royal Statist. Society Series B*, 65(3):679–700.

Carlin, B. and Chib, S. (1995). Bayesian model choice through Markov chain Monte Carlo. *J. Royal Statist. Society Series B*, 57(3):473–484.

Carlin, B., Gelfand, A., and Smith, A. (1992). Hierarchical Bayesian analysis of change point problems. *Applied Statistics (Series C)*, 41:389–405.

Carpenter, J., Clifford, P., and Fernhead, P. (1997). Building robust simulation-based filters for evolving datasets. Technical report, Department of Statistics, Oxford University.

Casella, G. and George, E. (1992). An introduction to Gibbs sampling. *American Statist.*, 46:167–174.

Casella, G., Lavine, M., and Robert, C. (2001). Explaining the perfect sampler. *American Statist.*, 55(4):299–305.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quater.*, 2:73–82.

Celeux, G. and Diebolt, J. (1990). Une version de type recuit simulé de l'algorithme EM. *Comptes Rendus Acad. Sciences Paris*, 310:119–124.

Chan, K. and Geyer, C. (1994). Discussion of "Markov chains for exploring posterior distribution". *Ann. Statist.*, 22:1747–1758.

Chen, M., Shao, Q., and Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

Churchill, G. (1995). Accurate restoration of DNA sequences (with discussion). In C. Gatsonis, J.S. Hodges, R. K. and Singpurwalla, N., editors, *Case Studies in Bayesian Statistics*, volume 2, pages 90–148. Springer–Verlag, New York.

Cressie, N. (1993). *Spatial Statistics*. John Wiley, New York.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Royal Statist. Society Series B*, 61(2):331–344.

Del Moral, P., Doucet, A., and Jasra, A. (2006). The sequential Monte Carlo samplers. *J. Royal Statist. Society Series B*, 68(3):411–436.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Society Series B*, 39:1–38.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Society Series B*, 56:363–375.

Dimakos, X. K. (2001). A guide to exact simulation. *International Statistical Review*, 69(1):27–48.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

Dupuis, J. (1995). Bayesian estimation of movement probabilities in open populations using hidden Markov chains. *Biometrika*, 82(4):761–772.

Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science, Special Issue*, pages 131–141. Available at `http://library.lanl.gov.cgi--bin/getfile?15--13.pdf`.

Erhman, J., Fosdick, L., and Handscomb, D. (1960). Computation of order parameters in an ising lattice by the Monte Carlo method. *J. Math. Phys.*, 1:547–558.

Fernández, R., Ferrari, P., and Garcia, N. L. (1999). Perfect simulation for interacting point processes, loss networks and Ising models. Technical report, Laboratoire Raphael Salem, Univ. de Rouen.

Fill, J. (1998a). An interruptible algorithm for exact sampling via Markov chains. *Ann. Applied Prob.*, 8:131–162.

Fill, J. (1998b). The move-to front rule: A case study for two perfect sampling algorithms. *Prob. Eng. Info. Sci.*, 8:131–162.

Fismen, M. (1998). Exact simulation using Markov chains. Technical Report 6/98, Institutt for Matematiske Fag, Oslo. Diploma-thesis.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Royal Statist. Society Series B*, 56:501–514.

Gelfand, A., Hills, S., Racine-Poon, A., and Smith, A. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. American Statist. Assoc.*, 85:972–982.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *J. American Statist. Assoc.*, 85:398–409.

Gelfand, A., Smith, A., and Lee, T. (1992). Bayesian analysis of constrained parameters and truncated data problems using Gibbs sampling. *J. American Statist. Assoc.*, 87:523–532.

Gelman, A. and King, G. (1990). Estimating the electoral consequences of legislative redistricting. *J. American Statist. Assoc.*, 85:274–282.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Science*, 7:457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *J. American Statist. Assoc.*, 88:881–889.

George, E. and Robert, C. (1992). Calculating Bayes estimates for capture-recapture models. *Biometrika*, 79(4):677–683.

Geyer, C. (1992). Practical Monte Carlo Markov chain (with discussion). *Statist. Science*, 7:473–511.

Geyer, C. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.*, 21:359–373.

Geyer, C. and Thompson, E. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. American Statist. Assoc.*, 90:909–920.

Gilks, W. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 4*, pages 641–649, Oxford. Oxford University Press.

Gilks, W. and Berzuini, C. (2001). Following a moving target–Monte Carlo inference for dynamic Bayesian models. *J. Royal Statist. Society Series B*, 63(1):127–146.

Gilks, W., Best, N., and Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statist. (Series C)*, 44:455–472.

Gilks, W., Roberts, G., and Sahu, S. (1998). Adaptive Markov chain Monte Carlo. *J. American Statist. Assoc.*, 93:1045–1054.

Gordon, N., Salmond, J., and Smith, A. (1993). A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing*, 140:107–113.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guihenneuc-Jouyaux, C. and Robert, C. (1998). Finite Markov chain convergence results and MCMC convergence assessments. *J. American Statist. Assoc.*, 93:1055–1067.

Hammersley, J. (1974). Discussion of Mr Besag's paper. *J. Royal Statist. Society Series B*, 36:230–231.

Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. John Wiley, New York.

Hammersley, J. and Morton, K. (1954). Poor man's Monte Carlo. *J. Royal Statist. Society Series B*, 16:23–38.

Handschin, J. and Mayne, D. (1969). Monte Carlo techniques to estimate the conditional expectation in multi-stage nonlinear filtering. *International Journal of Control*, 9:547–559.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57:97–109.

Hitchcock, D. B. (2003). A history of the Metropolis-Hastings algorithm. *American Statistician*, 57(4):254–257.

Hobert, J. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear models. *J. American Statist. Assoc.*, 91:1461–1473.

Hobert, J., Jones, G., Presnell, B., and Rosenthal, J. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, 89(4):731–743.

Hobert, J. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and px-da algorithms. *Ann. Statist.*, 36:532–554.

Jones, G., Haran, ., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. American Statist. Assoc.*, 101:1537–1547.

Jones, G. and Hobert, J. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Science*, 16(4):312–334.

Kemeny, J. and Snell, J. (1960). *Finite Markov Chains*. Van Nostrand, Princeton.

Kendall, W. and Møller, J. (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, 32:844–865.

Kipnis, C. and Varadhan, S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, 104:1–19.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non–Gaussian non–linear state space models. *J. Comput. Graph. Statist.*, 5:1–25.

Kong, A., Liu, J., and Wong, W. (1994). Sequential imputations and Bayesian missing data problems. *J. American Statist. Assoc.*, 89:278–288.

Kuhn, T. (1996). *The Structure of scientific Revolutions, Third Edition*. University of Chicago Press.

Landau, D. and Binder, K. (2005). *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, Cambridge, UK, second edition.

Lange, N., Carlin, B. P., and Gelfand, A. E. (1992). Hierarchal Bayes models for the progression of hiv infection using longitudinal cd4 t-cell numbers. *jasa*, 87:615–626.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals - a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214.

Liu, C. and Rubin, D. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotonous convergence. *Biometrika*, 81:633–648.

Liu, J. and Chen, R. (1995). Blind deconvolution via sequential imputations. *J. American Statist. Assoc.*, 90:567–576.

Liu, J., Wong, W., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes. *Biometrika*, 81:27–40.

Liu, J., Wong, W., and Kong, A. (1995). Correlation structure and convergence rate of the Gibbs sampler with various scans. *J. Royal Statist. Society Series B*, 57:157–169.

Liu, J. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *jasa*, 94:1264–1274.

Madras, N. and Slade, G. (1993). *The Self-Avoiding Random Walk*. Probability and its Applications. Birkhauser, Boston.

Marin, J.-M. and Robert, C. (2007). *Bayesian Core*. Springer-Verlag, New York.

Marshall, A. (1965). The use of multi-stage sampling schemes in Monte Carlo computations. In *Symposium on Monte Carlo Methods*. John Wiley, New York.

Meng, X. and Rubin, D. (1992). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278.

Meng, X. and van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86:301–320.

Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24:101–121.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*, 15:125–130.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *J. American Statist. Assoc.*, 44:335–341.

Møller, J. and Waagepetersen, R. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton.

Moussouris, J. (1974). Gibbs and Markov random systems with constraints. *J. Statist. Phys.*, 10:11–33.

Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *J. American Statist. Assoc.*, 90:233–241.

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–356.

Neal, R. (2003). Slice sampling (with discussion). *Ann. Statist.*, 31:705–767.

Pearl, J. (1987). Evidential reasoning using stochastic simulation in causal models. *Artificial Intelligence*, 32:247–257.

Peskun, P. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612.

Peskun, P. (1981). Guidelines for chosing the transition matrix in Monte Carlo methods using Markov chains. *Journal of Computational Physics*, 40:327–344.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252.

Qian, W. and Titterington, D. (1990). Parameter estimation for hidden Gibbs chains. *Statis. Prob. Letters*, 10:49–58.

Raftery, A. and Banfield, J. (1991). Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics. *Ann. Inst. Statist. Math.*, 43:32–43.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B*, 59:731–792.

Ripley, B. (1987). *Stochastic Simulation*. John Wiley, New York.

Robert, C. (1995). Convergence control techniques for MCMC algorithms. *Statis. Science*, 10(3):231–253.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition.

Roberts, G., Gelman, A., and Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Applied Prob.*, 7:110–120.

Roberts, G. and Rosenthal, J. (1999). Convergence of slice sampler Markov chains. *J. Royal Statist. Society Series B*, 61:643–660.

Roberts, G. and Rosenthal, J. (2005). Coupling and ergodicity of adaptive mcmc. *J. Applied Proba.*, 44:458–475.

Rosenbluth, M. and Rosenbluth, A. (1955). Monte Carlo calculation of the average extension of molecular chains. *J. Chemical Physics*, 23:356–359.

Rosenthal, J. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. American Statist. Assoc.*, 90:558–566.

Rubin, D. (1978). Multiple imputation in sample surveys: a phenomenological Bayesian approach to nonresponse. In U.S. Department of Commerce, S. S. A., editor, *Imputation and Editing of Faulty or Missing Survey Data*.

Smith, A. and Gelfand, A. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46:84–88.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Appl. Statist.*, 43:159–1789.

Stephens, D. A. and Smith, A. F. M. (1993). Bayesian inference in multipoint gene mapping. *Ann. Hum. Genetics*, 57:65–82.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.*, 28:40–74.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. American Statist. Assoc.*, 82:528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, 22:1701–1786.

Titterington, D. and Cox, D. (2001). *Biometrika: One Hundred Years*. Oxford University Press, Oxford, UK.

Wakefield, J., Smith, A., Racine-Poon, A., and Gelfand, A. (1994). Bayesian analysis of linear and non-linear population models using the Gibbs sampler. *Applied Statistics (Series C)*, 43:201–222.

Wang, C. S., Rutledge, J. J., and Gianola, D. (1993). Marginal inferences about variance-components in a mixed linear model using Gibbs sampling. *Gen. Sel. Evol.*, 25:41–62.

Wang, C. S., Rutledge, J. J., and Gianola, D. (1994). Bayesian analysis of mixed limear models via Gibbs sampling with an application to litter size in Iberian pigs. *Gen. Sel. Evol.*, 26:91–115.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. American Statist. Assoc.*, 85:699–704.

Zeger, S. and Karim, R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. American Statist. Assoc.*, 86:79–86.